JUSTSYSTEM

August 25, 1999

Richard L. Byrne, Esquire
Webb Ziesenheim Bruening Logsdon Orkin & Hanson
700 Koppers Building
Pittsburgh, Pennsylvania 15219-1818

RE: Invention Disclosure Record for "Method and Apparatus for Efficient Identification of Duplicate and Near-Duplicate Documents"
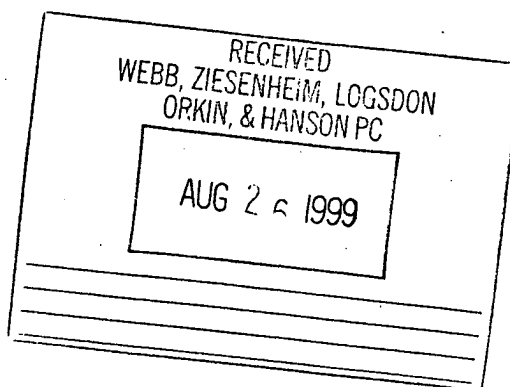
Dear Rick:

Enclosed please find a new Invention Disclosure Record from Mark Kantrowitz. Also enclosed is a floppy disk containing the source code for this invention.

We would appreciate it if you would review this information and advise us if you believe it will be worthwhile proceeding with a patent application. Thank you.

Sincerely,

Kimberly McDaniel
Business Manager

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815
FAX 412-471-4094
E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. McILVAINE III
BLYNN L SHIDELER

OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)

JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
DEBORAH M. ALTMAN
THOMAS J. CLINTON

PATENT AGENTS
DEAN E. GEIBEL
NATHAN J. PREPELKA

September 15, 1999

Kimberly McDaniel, Esq.
Business Manager
Justsystem Pittsburgh
    Research Center, Inc.
4616 Henry Street
Pittsburgh, PA 15213

> Re: Preliminary Patentability Search on "Method
> And Apparatus For Efficient Identification
> Of Duplicate And Near-Duplicate Documents And
> Text Spans Using High-Discriminability Text
> Fragments" Our File 991336

Dear Kim:

I want to acknowledge receipt of your August 25, 1999 letter including an Invention Disclosure Record and floppy disk for the above-identified invention. Please note our file number assigned to this matter.

We ordered copies of the three patents listed by Mark on page 9 of the invention disclosure. Can you send me a copy of pages 189-192 of the Data Structures book and a complete copy of the Patrick Juola article?

Thank you for sending us this disclosure for our review and possible search. We will report to you further after we have reviewed the patents and articles.

Very truly yours,

Richard L. Byrne

RLB/llm
cc: Kenneth G. Judson, Esq.

J

September 17, 1999

Richard L. Byrne, Esquire
Webb Ziesenheim Bruening Logsdon Orkin & Hanson
700 Koppers Building
Pittsburgh, Pennsylvania 15219-1818

RE: "Method and Apparatus for Efficient Identification of Duplicate and Near-Duplicate Documents"
   **Your File No. 991336**

Dear Rick:

Thank you for your letter dated September 15, 1999 regarding the above-referenced matter. Copies of the relevant pages from the <u>Data Structures</u> book and of the Patrick Juola article are enclosed.

Thank you for your assistance with regard to this matter. If you have any questions, or need any additional information, please do not hesitate to contact me.

Sincerely,

Kim

Kimberly McDaniel
Business Manager

J

JUSTSYSTEM
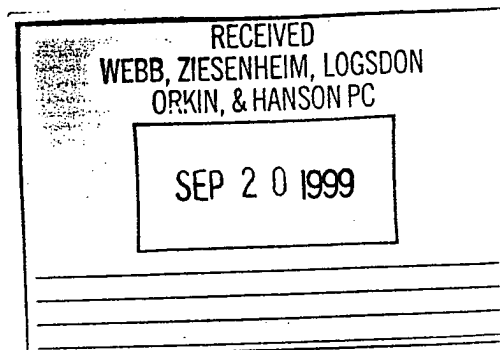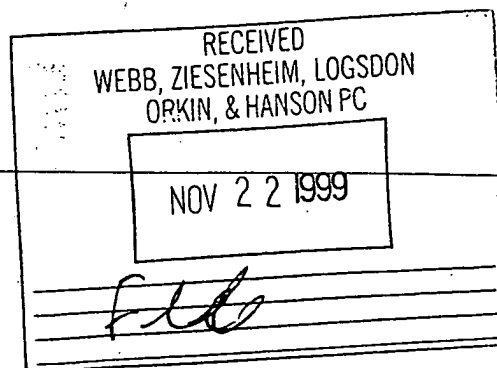
November 19, 1999

Richard L. Byrne, Esquire
Webb Ziesenheim Bruening Logsdon Orkin & Hanson
700 Koppers Building
Pittsburgh, Pennsylvania 15219-1818

RE: "Method and Apparatus for Efficient Identification of Duplicate and Near-Duplicate Documents"
    Your File No. 991336

Dear Rick:

Thank you for your letter dated November 12, 1999 with regard to the above-referenced matter. Scott Fahlman and Mark Kantrowitz have reviewed your letter and the patents enclosed with it, and we agree that you should proceed with the preparation of a patent application. Scott and Mark wanted to be sure that the application clearly identifies the difference between our invention and the previously-patented ones. Instead of trying to summarize their comments, I have enclosed their e-mail messages with this letter for you to review.

Thank you for all of your help. If you have any questions, or need any additional information, please do not hesitate to contact me or Mark at your convenience.

Sincerely,

Kimberly McDaniel
Business Manager

From: "Scott E. Fahlman" <sef@cs.cmu.edu>
To: mkant@jprc.com, mcdaniel@jprc.com
Subject: Patent application: Duplicate and near-duplicate documents
Date: Tue, 16 Nov 1999 15:52:19 -0500
Sender: Scott_Fahlman@CLYDE.BOLTZ.CS.CMU.EDU

It looks to me like the lawyers understand the essence of this invention,
and we should go forward with drafting the application as they recommend,
assuming Mark agrees.

The patents they dug up are not surprising or threatening. Two are brittle
checksum-based methods for detecting when whole documents are exact
duplicates of one another. We should make clear that our method is unique
in detecting near-duplicates. (I'm amazed that either patent was granted
-- this stuff is way beyond obvious, and equality of checksums has been
used for ages.)

The Queen patent is perhaps a bit more troublesome, since it looks for
exact matches of lines or sentences within the document, and uses these as
anchors for diff-like processing. Again, I'm amazed that this was granted,
but it's pretty old, so maybe pre-dates all of the obvious prior art.
Anyway, Mark should make clear exactly how we differ. I believe the key
differences are (a) focus on words and phrases, and (b) flexibility in
choosing which phrases to save and detect, based on how suitable they are
for the task of identifying near-duplicate documents. Specifically, we
want phrases that are unusual and characteristic of the document.

Or something like that...

-- Scott

X-Sender: mkant@mail.jprc.com
Date: Thu, 18 Nov 1999 15:51:51 -0500
To: "Scott E. Fahlman" <sef@cs.cmu.edu>,mcdaniel@jprc.com
From: Mark Kantrowitz <mkant@jprc.com>
Subject: Re: Patent application: Duplicate and near-duplicate documents
Cc: mkant@jprc.com

I agree with Scott. Please send these comments along with Scott's (or a
suitable summary, since I rambled a bit) to the patent attorneys. I think
they understand the essence of the invention very well and do not see any
reason why we shouldn't move forward with the drafting of the application.

I would suggest emphasizing a bit more the particular method of identifying
distinctive phrases (e.g., may not begin or end with a stop word, may
contain some of a particular class of stopwords within, restrictions on the
IDF values of the other words in the phrase), so that if the PTO tries to
disallow the more general claim of distinctive phrases on the basis of a
bad interpretation of Queen, in the worst case they will have to allow
claims for our specific method.

(Incidentally, the patents they included are the set I mentioned in the
prior art section of the invention disclosure. I go into a bit of detail
about the fragility of checksums there. I also show how, if one were to try
to generalize the state of the art such as LCS algorithm used by 'diff' to
identify near duplicates, it would still be inferior to our invention in
speed and accuracy.)

As a general rule, when I list a patent in an invention disclosure I only
do so after reviewing the entire text of the patent (but not the drawings).
The USPTO web site, http://patents.uspto.gov/, provides the full text (but
not the drawings) of all patents from January 1, 1976 to the present.

I agree that the Queen patent shouldn't have been granted from a computer
scientist's perspective, given the abundant prior art. (For example, the
GNU-Emacs command "compare-windows" existed in 1986.) Not the least reason
is that the algorithm as described is incomplete; they briefly mention the
case where one of the lines in the first file matches more than one line in
the second file, but their stated method of handling it will only work if
the two files are substantially the same (e.g., different versions of the
same document but for which the number of modifications is highly limited).
If you apply it to more divergent texts, it just doesn't work. Likewise,
their technique of addressing hash collisions is less likely to work for
more divergent texts. Of course, for the most common application their
technique for finding the differences between files will work well enough
to be useful. But it should be noted that such multiple matches for a line
is the principal reason for the LCS algorithm that's used in 'diff'.

But I suppose the patent examiners were either unaware of the prior art, or
thought their invention sufficiently different to be patentable.
Incidentally, the source-compare.lisp algorithm I included in the CMU AI
Repository uses a greedy approximation to 'diff' with limited lookahead
(e.g., work your way through the document until the first difference, and
compare lines after that to find the "closest" next matchup) has an
average-case linear running time, and so is more efficient in practice than
the Queen patent.

I'm not surprised that the USPTO granted their patent. I recently did a
search for "retractable leashes" because I thought I had a novel idea for a

better dog leash. All of the leashes that are commercially available
include a "stop" so you can artificially shorten the leash (e.g., for
walking on busy sidewalks). But this fixes the length, stopping the
retractability. The problem with this is if the dog runs back to you, the
leash lies on the ground and can get tangled in the dog's legs. It would be
better to have a retracting dog leash that allows you to set a limit on the
length of the leash while still allowing the retractability to work. While
doing a prior art search (turns out that someone else had the idea first,
but never incorporated it into commercial products), I found *many* pet
leash patents, all covering effectively the same claims. There are even
three different patents covering the combination of a flashlight with a
retractable dog leash, with overlapping claims. I'd love to be able to run
our duplicate detector on the USPTO patent database, since I'm sure we
would find many patents that are duplicates of one another.

Anyway, the differences Scott mentions between Queen and us:

  (a) we focus on words and phrases

  (b) we offer flexibility in choosing which phrases to save
      and detect, based on how suitable they are for the task
      of identifying near-duplicate documents

are but a few of the many differences (many of which I identified in the
invention disclosure), including:

 - Their invention identifies the differences between
   files, it doesn't identify when documents are
   duplicate or near duplicate.

 - Their invention cannot compare more than two documents.

 - Their invention requires the text to be in the same
   order in both files -- if I scramble the order of the
   sentences, as a student who is plagiarizing someone
   else's work might do, their invention will fall flat on
   its face.

 - My invention is more efficient than theirs.

 - Their invention requires long text spans (lines and sentences)
   because it won't work with shorter text spans. If you
   use shorter text spans, the number of collisions increases
   to the point where their algorithm is too slow to be
   practical. Although claim 12 mentions describes documents
   as having words, lines, and sentences, note that the claim
   omits the word "words" when referring to the hashing process.
   That's because their invention will not work efficiently or
   accurately with words or even phrases.

 - Their invention requires the lines to be the same in both
   documents. If you reformat the documents (e.g., change the
   linebreaks by changing the width of a page or even just
   deleting a single word or character), likely none of the
   lines will match. Even if they do the comparison on a
   sentence level, it will fail if the changes are widespread,
   such as a global text substitution (e.g., "i.e.," for "e.g.,",
   "this" for "that", "the" for "our", "Fahlman" for "Falhman").

Our invention is not subject to this kind of brittleness.

- Their invention does not provide any selectivity in the
  choice of anchorpoints, beyond requiring them to hash
  the same. My invention selects the more idiosyncratic
  phrases in common between documents, so that matches
  provide good evidence that the documents are related.
  As such, my invention has a set of criteria for choosing
  which text spans to use in comparing the documents. (If
  we don't have a claim covering this aspect of our invention,
  we should, since we're using less than the full set of
  text spans in common between the documents, and that is
  a new technique.)

So the bottom line is I think we have a clearly patentable invention which
includes several fairly broad patentable claims, and so we should proceed.

Mark
===

Mark Kantrowitz   mkant@just-research.com
Research Scientist   mkant@jprc.com
Justsystem Pittsburgh Research Center   http://www.jprc.com/users/mkant
4616 Henry Street   412-683-8674
Pittsburgh, PA 15213-3715, U.S.A.   412-683-4175 fax

JUSTSYSTEM

November 22, 1999

Richard L. Byrne, Esquire
Webb Ziesenheim Bruening Logsdon Orkin & Hanson
700 Koppers Building
Pittsburgh, Pennsylvania 15219-1818

RE: "Method and Apparatus for Efficient Identification of Duplicate and Near-Duplicate Documents"
        Your File No. 991336

Dear Rick:

Mark Kantrowitz has advised me that he just found a newly granted patent which might need to be cited in the above-referenced patent application. It is:

5,978,828 Greer, et. al. Intel. Nov. 2, 1999
URL bookmark update notification of page content or location changes

Mark said he initially thought that they were using the term "quotient" to mean a reduction of the web page to a simple value like a checksum. But upon deeper reading, it looks like it might be a way of associating version numbers and magnitude of changes with web pages. If the former, the arguments against checksums apply. If the latter, it is a mechanism for communicating that a page has changed but not for determining that a page has changed and the extent of the changes. So, if it's the latter, Mark doesn't think we need to cite it. But, if it's the former, we should probably cite it at the same place we cite the checksum patents.

If you have any questions, or need any additional information about this issue, please do not hesitate to contact me or Mark at your convenience.

Sincerely,

Kimberly McDaniel
Business Manager

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815
FAX 412-471-4094
E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

December 10, 1999

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. McILVAINE III
BLYNN L. SHIDELER

OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)
JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
DEBORAH M. ALTMAN
THOMAS J. CLINTON
DEAN E. GEIBEL
NATHAN J. PREPELKA
JESSICA M. SOSENKO

Kimberly McDaniel, Esq.
Business Manager
Justsystem Pittsburgh
    Research Center, Inc.
4616 Henry Street
Pittsburgh, PA 15213

> Re: Preliminary Patentability Search on "Method
>   And Apparatus For Efficient Identification
>   Of Duplicate And Near-Duplicate Documents And
>   Text Spans Using High-Discriminability Text
>   Fragments"   Our File 991336

Dear Kim:

        I want to acknowledge receipt of your November 19 and
November 22, 1999 letters regarding the above-identified invention.
Pursuant to your authorization, we are proceeding with the
preparation of a regular patent application on this invention. We
have assigned our file number 991842 to this application. We will
take U.S. Patent No. 5,978,828 into account when we prepare this
application.

        It is our understanding that this invention has not been
disclosed publicly or offered for sale. Therefore, we do not have
a firm date by which we must file the application in order to
preserve your United States and/or foreign patent rights. If your
plans have changed with respect to this invention, and you have
either disclosed the invention to the public or offered it for sale
in the past, or plan to do so in the near future, please let me
know so that we make sure that an application, even a provisional
patent application, is filed by the appropriate deadline.

        Thank you for entrusting this application to us. We will
have an application out to you for review by Mark, the sole
inventor, as soon as we can.

                        Very truly yours,

                        Richard L. Byrne

RLB/llm
cc: Kenneth G. Judson, Esq.

HARRISBURG OFFICE: 100 PINE STREET, HARRISBURG, PA 17108-1188
TELEPHONE 717-238-1555   FAX 717-238-1755

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815

FAX 412-471-4094

E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. McILVAINE III
BLYNN L. SHIDELER
———
OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)
JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
DEBORAH M. ALTMAN
THOMAS J. CLINTON
DEAN E. GEIBEL
NATHAN J. PREPELKA
JESSICA M. SOSENKO

February 11, 2000

VIA FEDERAL EXPRESS NO.: 8144 8840 9668

Kimberly McDaniel, Esq.
Business Manager
Justsystem Pittsburgh
     Research Center, Inc.
4616 Henry Street
Pittsburgh, PA 15213

> Re: **Mark Kantrowitz**
> United States Patent Application entitled
> "Method and Apparatus for Efficient Identification
> of Duplicate and Near-Duplicate Documents and Text
> Spans Using High-Discriminiability Text Fragments"
> Our File: 991842

Dear Kim:

Enclosed is a draft patent application for review by Mark Kantrowitz, the sole inventor. For your convenience, and to speed up the review process, I have enclosed one copy for you and Scott and one copy for Mark, who is not at the Justsystem Pittsburgh Research Center, Inc. address at the present time.

Please let me have the comments and changes from Mark and Scott, and we will prepare a final application for filing. If timing becomes an issue, we can file a revised application without formal papers from Mark and follow up with the formal papers after filing.
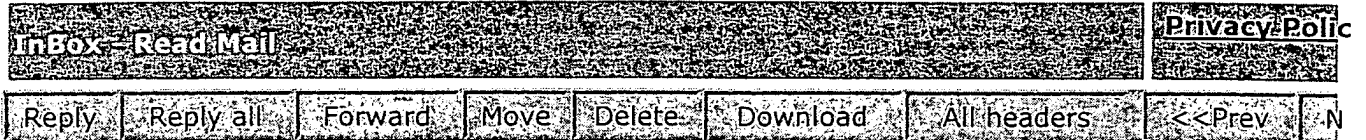
I look forward to hearing from you.

Very truly yours,

Richard L. Byrne

RLB/JMS/knw
Enclosures

cc: Kenneth G. Judson (w/o encs.)
        (Via First Class Mail)

Date: Thu, 22 Jun 2000 02:01:49 -0400
From: Scott Fahlman <Scott_Fahlman@myrddin.gwydion.cs.cmu.edu> Add to Address Book **Add Spam Block List**
Subject: Filing Mark's patents.
To: mcdaniel@jprc.com
Cc: mkant@jprc.com

Kim,

I've taken a quick look at the three draft applications Mark reviewed for
us, and I don't really have anything to add. So let's ask the Webb people
to make the changes Mark suggests and then file these ASAP. They may need
to contact Mark for clarification of his suggestions, so be sure they know
how to contact him while I'm away.

We'll need to get Mark's signature for each of these, though I guess this
can be after the filing if necessary. Note that he will be hard to reach
in August -- harder than usual.

Thanks,
Scott

==

3. Analyzing Affect (draft application). I have read the draft and did not
find any problems with it. I think we should file this application as soon
as possible.

==

4. Learning from user Self-Corrections (draft application). I recommend
making the following changes and corrections, after which it will be ready
for filing:

  + Page 1, line 16: Add something like "except when explicitly
    changed by the user (e.g., adding a word to a user dictionary)" to
    the end of the sentence. There are rule-based systems that allow the
    user to add rules, but the user has to explicitly devise and add the
    rule, instead of having the system implicitly infer and
    automatically add the rule. The paragraph is fine as is, but we
    might want to add that clause to clarify.

  + Page 18, line 5: Replace "the C language" with "the Lisp
    language". Line 14: Replace "not really part of" with "not normally
    applicable to".

  + Page 28, claim 1: The first step listed, "changing current
    text into transformed text" is not a step performed by the computer,
    but by the user, while the other two steps are performed by the
    computer. Perhaps it's ok for the claim to mix steps performed by
    the user and the computer (it is, after all, a "computer-assisted
    method"), but it might be better to clarify, perhaps by adding

"observing the user" to the beginning of that step, so that all three steps related to the computer's actions.

+ Pages 28-36: I noticed the number "5" appearing in the left margin on my copy on these pages. It looks like a leftover from the line numbering.

+ Claim 17, 18: The phrase "the at least one other rule" sounds a bit awkward, but that might be the style with which such claims are written.

+ Claim 19, 21, 23: The phrasing of these claims is a bit awkward. All are of the form: "further including the step of if x, y". Perhaps the "if x, y" should be set off on a line by itself, or the clauses inverted, to read "y, if x".

==

5. Efficient Identification of Near Duplicate Documents (draft application). I recommend making the following changes and corrections, after which it will be reading for filing:

+ Page 1, lines 21-22: Delete the extraneous carriage return within the word "another".

+ Page 4, lines 17+. I think the draft doesn't go into enough detail about distinctive features. The key to the present invention is a method for a priori identification of likely distinctive text fragments. This paragraph is rather unspecific. For example, where it says on line 19-20 "word n-grams", that is a bit too general. We do NOT count all word n-grams, but rather only a specific subset that are distinctive for the document and likely to appear in duplicate and near duplicate documents. This means, for instance, that the individual words within the n-gram cannot be so rare as to be unique to the document. So we use words that are not as rare, and hence found in several documents, but achieve distinctiveness by using n-grams containing those words. Another novel aspect is the inclusion of "glue words" (very common words, such as words normally treated as stopwords) within the distinctive n-gram (but not at either end). Our distinctive phrases can include words at either extrema (words that are common to just a few documents, and/or words that are common to all but a few documents), but not words of intermediate rarity.

Basically, it looks like the "Summary of Invention" section skips over the discussion of distinctive features a bit too quickly. It does a wonderful job of discussing all the various applications, and an adequate job of showing how distinctive features may be used to efficiently compare a large collection of documents (another key innovation of the invention), but is a bit thin on the discussion of how distinctive features are obtained. (The in-depth discussion later is sufficiently detailed, as are the relevant claims.)

Admittedly, the invention is the first to use distinctive features in this fashion to identify near-duplicate documents, and so the most valuable aspect of the patent is in the overall method, not the specific type of distinctive features. Yet we should elaborate a bit on the nature of the distinctive features and their properties (e.g., must not be common to just one document, must not be common to a large number of documents, etc.).

+ The extensions of the method to dealing with images are not a key aspect of the application, and may represent a distraction. I would not object if you decided to delete them. If so, you will need to delete the clause beginning "and detecting copyright infringement

of images" on lines 26-28 of page 5, and claims 39-43

<< I agree that these claims should be dropped, in order to maximize the >>
<< focus on the more valuable and better-developed text-based part of the >>
<< invention. That's what we most want to protect. -- Scott >>

    +   Page 16, line 21: insert
            "United States of America" yields
        before
            "United States"
        and
            upon splitting at the "of"
        after.

    +   Page 18, line 4: Replace "a copy" with "the original document".

===

```
[ Reply ] [ Reply all ] [ Forward ] [ Move ] [ Delete ] [ Download ] [ All headers ] [ <<Prev ] [ N
```

JUSTSYSTEM
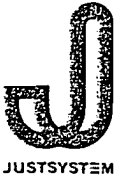
June 27, 2000

Richard L. Byrne, Esquire
Webb Ziesenheim Bruening Logsdon Orkin & Hanson
700 Koppers Building
Pittsburgh, Pennsylvania 15219-1818
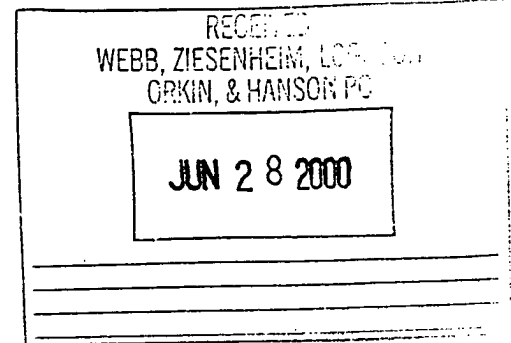
Dear Rick:

Enclosed are the inventors' comments on draft applications for six of the files you have for us. As I mentioned to you on the telephone, I don't have my files any more, so I can't provide you with file numbers. If you have any trouble matching up the comments with your files, let me know and I'll see what we can figure out (sometimes we call things by a slightly different title than you do!).

Scott mentions in his attached comments that Mark Kantrowitz may be difficult to reach, and that Rahul was out of the country. Rahul is now back in the U.S., although he is no longer in the Pittsburgh area. The best way to get in touch with both Mark and Rahul is probably by email. Their addresses are mkant@jprc.com and rahuls@cs.cmu.edu.

Thank you.

Sincerely,

Kimberly McDaniel
Business Manager

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815
FAX 412-471-4094
E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. MCILVAINE III
BLYNN L SHIDELER

OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)

JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
THOMAS J. CLINTON
DEAN E. GEIBEL
NATHAN J. PREPELKA
JESSICA M. SOSENKO
KIRK M. MILES

PATENT AGENTS
J. MATTHEW PRITCHARD
GARY F. MATZ

October 10, 2000

Justsystem Corporation
c/o Scott E. Fahlman
5259 Fair Oaks Street
Pittsburgh, PA 15217

Re: Proposed U.S. Patent Application entitled "Method
and Apparatus for Learning from User Self-Corrections,
Revisions and Modifications"
Our File: 991601

Proposed U.S. Patent Application entitled "Method and
Apparatus for Analyzing Affect and Emotion Text"
Our File: 991710

Proposed U.S. Patent Application entitled "Method and
Apparatus for Efficient Identification of Duplicate and Near-Duplicate
Documents and Text Spans Using High-Discriminability Text Fragments"
Our File: 991842

Dear Scott:

We have finalized the above-identified applications in view of the inventors' comments. Enclosed are final versions of the applications. Each application has attached at the end a document entitled "Declaration and Power of Attorney". Also enclosed for each application is a separate "Assignment".

Mark Kantrowitz is a sole inventor on two of the applications and is a joint inventor on the third application. Mark should review each application for accuracy and completeness. Ray Pelletier and Evan Bernstein should review for accuracy and completeness only the application that names them as inventors. If the applications are satisfactory, have Mark alone or with the other inventors, as appropriate, sign and date in blue ink, each formal document where indicated. Since each Assignment refers to the application "for which we have this day

executed an application for United States Letters Patent", each inventor should sign the appropriate Assignment on the same day he signs the Declaration and Power of Attorney attached to the application. The signatures on the Assignments should be notarized if at all possible.

Alternatively, I can file the enclosed applications without their signatures and without paying any filing fees at this time. We can follow up later, at a more leisurely pace, with formal papers that specifically identify the serial numbers and filing dates for these applications. This option will speed up the filing process, but will require the payment of a $130.00 surcharge per application for submitting the formal papers at a later date. I hesitate to follow this route since Mark may well have additional changes.

The government fees for filing these applications (not counting the $130.00 surcharge if we follow the option discussed above) will be as follows:

File No.	Filing Fee	Assignment Fee	TOTAL
991601	$1,178.00	$40.00	$1,218.00
991710	836.00	40.00	876.00
991842	1,052.00	40.00	1,092.00

Please arrange to have these fees transferred by Justsystem Corporation to our account. If we follow the alternative filing route discussed above, increase the filing fee for each application by $130.00.

We are not facing any deadline for filing these applications, but it would be advisable to file them as quickly as possible. Let me know if you have any questions at this time. Otherwise, I look forward to receiving the signed application papers and wire transfer as discussed above.

Very truly yours,

Richard L. Byrne

RLB/JMS/knw
Enclosures

cc: Kenneth G. Judson (w/o encs.)

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815
FAX 412-471-4094
E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. MCILVAINE III
BLYNN L. SHIDELER

OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)

JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
THOMAS J. CLINTON
DEAN E. GEIBEL
NATHAN J. PREPELKA
JESSICA M. SOSENKO
KIRK M. MILES

PATENT AGENTS
J. MATTHEW PRITCHARD
GARY F. MATZ

October 13, 2000

Justsystem Corporation
c/o Scott E. Fahlman
5259 Fair Oaks Street
Pittsburgh, PA 15217

Re: Proposed U.S. Patent Application entitled "Method
and Apparatus for Learning from User Self-Corrections,
Revisions and Modifications"
Our File: 991601

Proposed U.S. Patent Application entitled "Method and
Apparatus for Analyzing Affect and Emotion Text"
Our File: 991710

Proposed U.S. Patent Application entitled "Method and
Apparatus for Efficient Identification of Duplicate and Near-Duplicate
Documents and Text Spans Using High-Discriminability Text Fragments"
Our File: 991842

Proposed U.S. Patent Application entitled "Method and Apparatus for
Vision-Based Coupling Between Pointer Actions and Projected Images"
Our File: 000746

Dear Scott:

Per our recent e-mails, enclosed are the Declarations with Mark Kantrowitz's correct address. Please have the papers signed by the appropriate inventors and return to me with the applications and signed Assignments at your earliest convenience.

Very truly yours,

Jessica M. Sosenko

JMS/knw
Enclosures

cc: Kenneth G. Judson (w/o encs.)

From:	Scott E. Fahlman [sef@cs.cmu.edu]
Sent:	Friday, October 27, 2000 6:47 AM
To:	rbyrne@webblaw.com
Subject:	Kantrowitz patent application, Webb # 991451

Rick,

I've spent the last couple of days digging through the stacks of
patent-related mail and documents. I think I've now got everything
under
control. My biggest to-do item at present is tracking down the
inventors
for all the final-draft applications to get their notarized signatures.
I
also have asked Japan for instructions about the foreign filing
decisions
that must be made soon.

I came across your letter of September 1, 2000, which contained good
news
about the Kantrowitz TLTF patent, Webb file number 991451. This letter
contained the welcome news that the PCT examination was in our favor.

You recommend abandoning our current USPTO application and instead
pursuing
U.S. protection under the "national phase" of the PCT process. I don't
really understand the advantage of doing it this way, but if you believe
this will improve our odds of getting a timely and positive decision, I
will trust your judgement in this. Do what you think is best in this
case.
The additional $100 fee is not a problem if you prefer to go this way.

Best regards,
Scott

```
=====================================================================
===
```
Scott E. Fahlman   Internet: sef@cs.cmu.edu
Principal Research Scientist   Phone: 412 268-2575
Department of Computer Science   Fax: 412 268-5576
Carnegie Mellon University   Latitude: 40:26:46 N
5000 Forbes Avenue   Longitude: 79:56:55 W
Pittsburgh, PA 15213   Mood: :-)
```
=====================================================================
===
```

WILLIAM H. LOGSDON
RUSSELL D. ORKIN
DAVID C. HANSON
FREDERICK B. ZIESENHEIM
RICHARD L. BYRNE
KENT E. BALDAUF
BARBARA E. JOHNSON
PAUL M. REZNICK
JOHN W. MCILVAINE III
BLYNN L. SHIDELER

OF COUNSEL
MICHAEL I. SHAMOS

WILLIAM H. WEBB (1929-1997)

WEBB ZIESENHEIM LOGSDON ORKIN & HANSON, P.C.
700 KOPPERS BUILDING
436 SEVENTH AVENUE
PITTSBURGH, PA 15219-1818
TELEPHONE 412-471-8815
FAX 412-471-4094
E-MAIL webblaw@webblaw.com

PATENT, TRADEMARK & COPYRIGHT LAW

FAX TRANSMITTAL
PAGE 1 of 2
November 9, 2000

JULIE W. MEDER
LESTER N. FORTNEY
RANDALL A. NOTZEN
JESSE A. HIRSHMAN
JAMES G. PORCELLI
KENT E. BALDAUF, JR.
CHRISTIAN E. SCHUSTER
THOMAS J. CLINTON
DEAN E. GEIBEL
NATHAN J. PREPELKA
JESSICA M. SOSENKO
KIRK M. MILES

PATENT AGENTS
J. MATTHEW PRITCHARD
GARY F. MATZ

FACSIMILE NO.: 011-81-88-666-1170

Mr. Hidefumi Hata
Justsystem Corporation
Brains Park
Tokushima 771-0189
JAPAN

> Re: KANTROWITZ U.S. Patent Application entitled "Method And
> Apparatus For Efficient Identification Of Duplicate And Near-
> Duplicate Documents And Text Spans Using High-Discriminability
> Text Fragments"   Our File 991842

Dear Mr. Hata:

Scott Fahlman authorized us to file the above-identified application on behalf of Justsystem Corporation and we will do so in the near future.

Filing the application will require our payment of $1,092.00 in government filing fees. Our invoice is attached for that payment. Please arrange to have this amount wire transferred as soon as possible to our account identified in the invoice. We will not hold up filing the application while awaiting your payment, but we would appreciate receiving this payment promptly to cover our out-of-pocket expense.

Thank you for your attention to this matter.

Very truly yours,

Richard L. Byrne

RLB/llm
Attachment
cc: Mr. Scott E. Fahlman (w/attachment via first class mail)

November 15, 2000

Justsystem Corporation
c/o Mr. Scott E. Fahlman
5259 Fair Oaks Street
Pittsburgh, PA 15217

Re: U.S. Patent Application entitled "Method And Apparatus For
      Analyzing Affect And Emotion In Text"   Our File 991710

      U.S. Patent Application entitled "Method And Apparatus For
      Efficient Identification Of Duplicate And Near-Duplicate
      Documents And Text Spans Using High-Discriminability
      Text Fragments"   Our File 991842

Dear Scott:

I want to let you know that the above-identified patent applications were mailed to the United States Patent and Trademark Office for filing on November 15, 2000 by Express Mail. By using Express Mail service, the applications should be given November 15, 2000 as their official filing dates.

We will advise you in due course of the official serial numbers and filing dates assigned to these applications. At that time, we will send you and Mr. Shiozaki a complete copy of the application papers as filed.

Very truly yours,

Richard L. Byrne

RLB/llm
cc: Mr. Kenya Shiozaki